Tsunami Early Warning Detection using Bayesian Classifier

by Dewi Yanti Liliana

2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)

Tsunami Early Warning Detection using Bayesian Classifier

Dewi Yanti Liliana
Department of Informatics and Computer Engineering
Politeknik Negeri Jakarta
Depok, Indonesia
dewiyanti.liliana@tik.pnj.ac.id

Diah Priharsari
Faculty of Engineering and Information Technology
University of Technology Sydney
Sydney, Australia
diah.priharsari@students.uts.edu.au

***Abstract*— Tsunami is a term for disaster where the seawater rises to the land caused by the great earthquake with a shallow epicenter in the ocean. Indonesia is vulnerable to the tsunami; especially in the area where the junction of Eurasia, Indo-Australia, and Pacific Plates meet. Soon after the earthquake happened, there is a time interval before a tsunami to come. The break-time can be used to alert people to evacuate themselves from the Tsunami. This study proposed a tsunami early warning system, which autonomously predicts the potential of tsunami hazard using machine learning techniques. Bayesian classifier was trained to predict the tsunami potential. The tsunami training data was taken from the InaTews website; a national project of Indonesia that involves many institutions, local or international. InaTews informed the real-time data of earthquakes and tsunami that happened in Indonesia. The proposed methods extracted some invariant characteristics from the data and trained the machine learning classifiers to predict the potential result; either tsunami or not-tsunami using three parameters of earthquake: magnitude, epicenter, and location. The experiment gave an optimistic result; using a varying number of training data, it gained 92.37% on the average accuracy rate and 0.98 on the F1 score.**

Keywords*— *bayesian probabilistic classifier, early warning system, support vector machine, supervised machine learning, tsunami mitigation

I. INTRODUCTION

"Tsunami" term comes from the Japanese language. It describes a situation where an enormous sea wave reaches the land, which may be caused by the great earthquake with a shallow epicenter in the ocean. The geographical location of Indonesia is traversed by the confluence of three tectonic plates, namely: Indo-Australian Plate, Eurasian plate and the Pacific plate. These plates meet at the seafloor; hence, if a major earthquake happens with a shallow depth; it potentially causes a tsunami [1]. Fig. 1 shows a map of Indonesian potential areas highlighted in red line which are vulnerable to earthquakes and tsunami.

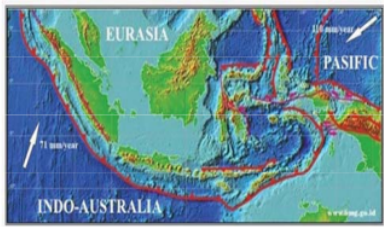

Fig. 1. The map of Indonesian potential tsunami areas

A massive earthquake with tsunami disasters that happened in Aceh in 2004 and Palu in 2018 has resulted in a hundred thousand of victims and property losses. Mitigation efforts are required by both national governments as well as the community to reduce the risk of catastrophic earthquakes and tsunami [2]. One of the mitigation efforts is building the tsunami early warning system to prevent the risk of higher losses. Tsunami Early Warning System is a series of complex systems that involve many stakeholders at international, regional, national, and community levels [3].

The mitigation action by the Indonesia government is building the Indonesian Tsunami Early Warning System, or in short InaTews [4]. InaTews is a national project involving various institutions in the nation under the coordination of the Ministry of Research and Technology. InaTews gains the earthquake data recorded by seismic instruments. The data is sent to the Tsunami Warning Centre and analyzed by the InaTews system. The result is classified based on the tsunami potential. The analytical method used for classification of tsunami is a Decision Support System (DSS) [4]. This result is immediately dispersed to the government agencies who are in charge of informing the general public through the warning sirens, TV, radio, and cable television [3]. Although seismographic data analysis has been conducted using the DSS, but it still needs improvement.

Earthquakes are caused by the sudden release of the earth's energy, being marked by the breaking of rock layers in the earth's crust. The accumulation of the earthquake energy is generated from the movement of tectonic plates. This energy emitted to all directions in the forms of an earthquake wave, so its effects can be felt up to the surface of the earth [2]. The results of the earthquake are ground shaking, liquefaction, soil avalanches, and Tsunami. Earthquake parameters include the time of the earthquake, the location of the epicenter, epicenter depth, and the strength of the earthquake (magnitude) [2].

The true meaning of tsunami is the vertical displacement of water bodies caused by changes in sea surface [5]. Changes in a sea-level may occur due to the earthquake centered under the sea, underwater volcanic eruptions, underwater landslides, or a meteor hit down the sea. Tsunami waves can spread in all directions. At sea, a tsunami can travel at the speed of 500 - 1000 km/h, equivalent to the speed of the airplane. The wave height in the ocean is probably about one meter, thus the rate of the wave may not be felt by the sailing ship on the sea. When approaching the shore, the wave speed will drop to 30 km/h, but the height has been increased up to tens of meters. Lacing tsunami waves can go up to tens of kilometers from shore. Damage and loss are caused by the blow of water and materials carried by tsunami waves [6].

The earthquake which can generate a tsunami has several characteristics: the epicenter is located under the sea; the depth of the epicenter is relatively shallow, less than 70 km; the magnitude (M) is large (M> 7.0 Richter scale) [2].

Indonesia is vulnerable to tsunamis, particularly the islands which directly border with the Eurasian plate, the Indo-Australian and the Pacific plate, including the Western part of Sumatra Island, South Java Island, Timor Island, Northern Papua Island, Sulawesi, and Maluku, and East Part of Borneo Island.

Previous studies determined the tsunami hazard rating by using Analytical Hierarchy Process (AHP) [7]. Another research ranked the tsunami hazard by using the data mining C 4.5 algorithm [8]. However, the research has lacked information about tsunami potential caused by the earthquake. Dilectin and Mercy implemented Tsunami classification using KNN for real-time data, and automatic detection of novel class for the outliers [9]. Meanwhile, the computer vision technique was used to support the Tsunami Warning System by extracting the tsunami-like wave video [10]. Wachter, et al. [11] also emphasized the future challenge of tsunami early warning system, which needed more improvement, including using a machine learning algorithm to predict the tsunami potentials.

In this paper, we developed a tsunami early warning system which autonomously analyzed tsunami potential using a reliable method of machine learning; Bayesian Probabilistic Supervised Learning or Bayesian classifier. Three parameters of an earthquake are used: magnitude, epicenter, and location to predict the occurrence of a tsunami caused by an earthquake. The training data was obtained from the InaTEWS website [4]. It is expected that this system can be integrated with a seismic recording device so that the classification can be done in real-time.

II. MATERIALS AND METHODS

Probabilistic Supervised Learning is one of the machine learning methods based on probabilistic statistical theory. In supervised learning, there exists a set of training data that has been labeled with categories or weights for each pattern [12]. One of the Probabilistic Supervised Learning techniques for pattern recognition is Bayesian Probabilistic Classifier. It is a reliable yet simple machine learning techniques to recognize patterns.

Bayesian Probabilistic Classifier or so-called Bayesian Classifier utilizes conditional probability theory. It predicts the probability of the future event based on the past experience. Many applications implement this algorithm. Two groups of researchers Pantel and Lin, and Microsoft Research, introduced a Bayesian statistical method for anti-spam filters [13]. Some studies were conducted on tsunami hazard applications [14-15].

A. Bayesian Probabilistic Classifier for Tsunami Classification

Let say X represents a set of attributes, and Y represents a class variable. If the class variable has a non-deterministic relationship with attributes, then X *and* Y can be treated as random variables that have a conditional probability, denoted by $P(Y|X)$. This conditional probability is also known as the posterior probability of Y given prior probability $P(Y)$. During the training phase, it is necessary to study the posterior probability for the entire combination of X and Y based on the information obtained from the training data. By knowing these opportunities, test data X' can be classified by finding the Y' class that maximizes the posterior probability $P(Y|X)$.

Bayes Theorem provides a term of posterior probability $P(Y|X)$ from the prior probability $P(Y)$, conditional class probability or so-called likelihood $P(X|Y)$, and evidence $P(X)$. Eq. 1 formulates the Bayesian Theorem.

$$P(Y|X)=\frac{P(X|Y).P(Y)}{P(X)} \quad (1)$$

Bayesian Probabilistic Classifier estimates the conditional probability of a class by assuming that attributes of a class have conditional independent characteristics. Conditional independent assumption can be expressed in (2).

$$P(X|Y=y)=\prod_{i=1}^{d}P(X_i|Y=y) \quad (2)$$

Each set of data $X=\{X_1, X_2, \ldots, X_d\}$ consists of d attributes. Let X, Y, and Z represent three sets of random variables, X conditionally Y is stated to be conditionally independent given Z. It can be expressed in (3).

$$P(X,Y|Z)=P(X|Z).P(Y|Z) \quad (3)$$

To classify the test data, Bayesian classifier computes the posterior probability of each class in Y using (4). Since $P(X)$ is constant for every Y, it can be ignored.

$$P(Y|X)=\frac{P(Y)\prod P(X_i|Y)}{P(X)} \quad (4)$$

For a discrete attribute, conditional probability is estimated based on the occurrence of X *in* class Y. While for continuous attribute, conditional probability is assumed as the distributions of the training data. Gaussian distribution is often chosen to represent the class conditional probability for continuous attributes. The distribution is characterized by two parameters, namely, mean μ and variance σ^2. For each class y_i, the class conditional probability for attributes x_i is:

$$P(X_i=x_i|Y=y_j)=\frac{1}{\sqrt{2\mu\sigma_{ij}}}exp^{\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad (5)$$

Parameters μ_{ij} can be estimated based on the sample mean of entire training data with label class y_j. In the same way, the sample variance can be estimated from the same training data.

B. Design of Tsunami Classifier

Five major steps in machine learning and pattern recognition are including data collection, feature selection, model selection, classifier training, and evaluation. Fig. 2 depicts these stages. Data collection is an important initial step of the supervised learning method. At this stage, representative data is collected, the test data is also collected for classifier testing purpose. The data source is obtained from the geophysical data from the website http://inatews.bmkg.go.id. Feature selection is a process to determine the main characteristics of the data that form two classes, tsunami, and not-tsunami. Model selection is the adjustment of Bayesian Classifier parameters for tsunami detection. Classifier training is feeding the model with the labeled training data so that the classifier can learn to model the tsunami prediction. Hence, the performance of the classifier can be evaluated using evaluation tools such as accuracy, precision, and recall. We validate the model by comparing the result with the benchmark data from InaTews.

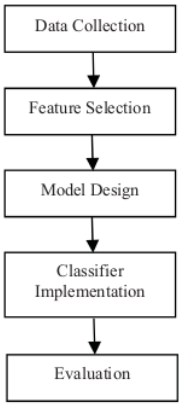

Fig. 2. The classification methodology

We collected hundreds of seismic data from the InaTews website in July 2019, exactly 350 training data and 320 testing data. Raw data was a webpage consist of specific information on earthquake occurrence, location point, and details of earthquake and tsunami potential. The preprocessing step converted the raw data into a flat-file with meaningful information for the tsunami classification.

On the feature selection stage, we analyze the determined characteristics of the earthquake, which may cause a tsunami. The features must be invariant to distinguish between two possible categories; earthquake with tsunami potential, and earthquake with no tsunami potential. We proposed three main features of the earthquake that influenced tsunami; the epicenter depth, the magnitude, and the location of the earthquake. Fig. 3 gives an example of the raw data from InaTews website, written originally in *bahasa* language. It gives information about three tsunami features: 36 Km in depth; 7.0 SR in magnitude; and the location took place in the ocean. The depth and magnitude are continuous features (numeric value), while the location is discrete feature (land or ocean).

Magnitudo	7,0
Tanggal & Waktu Kejadian	07-Jul-2019 22:08:42 WIB
Lokasi	0.54 LU - 126.19 BT
Kedalaman	36 Km
Keterangan Lokasi Gempabumi	133 km BaratDaya TERNATE-MALUT 134 km BaratDaya TERNATE-MALUT 140 km BaratDaya TIDORE-MALUT 153 km BaratDaya JAILOLO-MALUT 2277 km TimurLaut JAKARTA-INDONESIA
Potensi	Potensi TSUNAMI utk dtrskn pd msyrkt

Fig. 3. An example of raw data taken from the InaTEWS

The next stage is the model design. At this stage, we designed and formulated the Bayesian Probabilistic Classifier based on (4) that corresponded to the prediction of tsunami potential. Eq. 4 can be used as the kernel function in a probabilistic-based classification problem. Evidence *P(X)* can be omitted without changing the meaning. The formula for the classification of two classes, tsunami and not a tsunami, are shown on (6) and (7).

$$P(Y_1|X) = P(Y_1).P(X_1|Y_1).P(X_2|Y_1).P(X_3|Y_1) \quad (6)$$
$$P(Y_2|X) = P(Y_2).P(X_1|Y_2).P(X_2|Y_2).P(X_3|Y_2) \quad (7)$$

where Y_1 denotes tsunami, and Y_2 denotes not tsunami class categories. X_1, X_2, X_3 denote three features: depth, magnitude, and location, respectively. Eq. 6 computes the posterior probability of tsunami class $P(Y_1|X)$, and (7) computes the posterior probability of not-tsunami class $P(Y_2|X)$. While likelihood $P(X_1|\ Yi)$ and $P(X_2\ |\ Yi)$ are continuous features which are obtained from equation 5, and $P(X_3\ |\ Yi)$ is a discrete feature that is computed from the occurrence of X_3 within the entire training data. The decision is made by comparing the value of both posteriors. The largest posterior is selected as the prediction result, as in (8).

$$P(Y_{Tsunami}|X) > P(Y_{NotTsunami}|X) \quad (8)$$

The next stage is the classifier implementation, where the model that has been designed is implemented. Two phases of supervised learning are the training phase and the testing phase. At the training phase, we provided a set of training data to the classifier so that it can learn and build a classifier model. At the testing phase, a set of testing data is provided. The classifier has learned from the data in the previous phase. Parameters are given as well, including the prior probability of each class, mean and variance of the continuous feature of the training data, and the probability of each discrete feature in training data.

The last stage is the evaluation. Evaluation of system performance is done by measuring the precision and recall of the classifier performances by giving a variety size of training data. At the evaluation stage, we can notice the effect of training data size on the accuracy of classification. We also test the classifier with several numbers of testing data. The results will be compared with the benchmark data (real result from InaTews source data), and it is used to determine the performance of the classifier.

III. Results and Discussion

As a supervised learning method, the training data hold an important part in the Bayesian classification process. The data was taken from Indonesia Tsunami Early Warning System or InaTews website (http://inatews.bmkg.go.id/). The number of training data was 350, and the testing data was 320. Several experiments using a different number of training data were performed. The number of training data increased gradually during the experiment. The results of the experiment described the effect of the number of training data to the classifier performance. We observed the F1 score, which combined precision and recalls vales. F1 score is a common analysis tool on pattern recognition, the value is between 0 -1; the greater the value, the better the performance. F1 score is calculated using (9).

$$F1 = 2(precision * recall)/(precision + recall) \quad (9)$$

We also present the accuracy rate of the classifier. The average accuracy rate (AAR) for all the testing scenarios using a different number of the training set is 92.37%. It represents a powerful capability of Bayesian classifier for the tsunami classification task. Table I describes the results of the experiment. The number of training data is varying to provide the model with a different variation of training data numbers.

TABLE I. AVERAGE ACCURACY RATE (ARR)

# Training data	num. of testing data				
	30	40	50	60	70
10	0.29	0.43	0.83	0.43	0.39
20	0.67	0.83	0.83	0.80	0.73
30	0.67	0.80	0.67	0.74	0.70
40	0.80	0.89	0.83	0.88	0.82
50	0.80	1.00	0.83	0.94	0.89
60	0.80	1.00	0.71	0.94	0.89
70	0.80	1.00	0.83	0.94	0.89
AAR (%)	91.43	93.57	94.57	91.67	90.61

In the second scenario, we conducted an experiment using fixed-size training data with different sizes of testing data (from 30 to 70 data) to see the classification result. The result can be seen in Table II.

TABLE II. THE F1 SCORE

#Test data	TP	FP	FN	TN	Precision	Recall	F1
30	27	0	1	2	1	0.96	0.98
40	35	0	1	4	1	0.97	0.99
50	43	0	2	5	1	0.96	0.98
60	51	0	2	7	1	0.96	0.98
70	60	0	3	7	1	0.95	0.98

We can read from Table II, which is using 40 numbers of training data, the classifier has a high precision rate of 1. There is no false-positive alarm in it. It means that the classifier does not misclassify any "tsunami" testing data into the "not-tsunami" class. In other words, the classifier has been successfully predicted the "tsunami" testing data as it is. Meanwhile, the recall rate is 0.96 on average. Some misclassified predictions indicated by the false negative alarm represents the false classification of some "not- tsunami" testing data were classified as "tsunami" class. These results can be explained due to the composition of the training data. Since we all know that the occurrence of a tsunami disaster is somehow very rare, the training data consists of imbalance class categories. Lack of the number of tsunami class data; in the opposite, significant in the number of not-tsunami class data. The classifier supposed to learn a wide variety of class categories so that it can enhance its prediction capability.

Fig. 4 depicts the F1 score for the experiment in Table II. The maximum value of F-measure is 1. It indicates the high precision and recall rate. The experiment which uses 40 training data yields an auspicious result, where values of F1 are high for all testing data, and the average F1 score is 0.98. We also compare the F1 score of our proposed method with other machine learning methods using Support Vector Machine (SVM) as the classifier. Under the same training-testing data, SVM got 0.97 of F1 score. Meanwhile, our proposed Bayesian classifier got 0.98 of the average F1 score. Thus, it implies that the proposed method demonstrates a high classification performance.

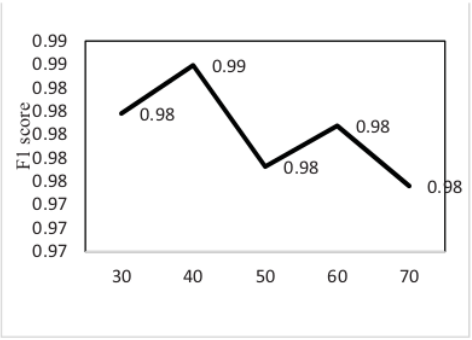

Fig. 4. Graphic of F1 score using 40 testing data

IV. CONCLUSION

The output of the proposed model is tsunami prediction from the earthquake event. Based on the implementation results, the Bayesian Probabilistic Classifier for tsunami early warning system has revealed high-quality classification results, indicated by the high accuracy and F1 score. To some extent, this research is dealing with imbalanced training data, which provides less information about the occurrence of some classes. Since the occurrence of tsunami is rare to compare to that of not-tsunami events. In supervised learning, we need to provide a good quality training data; balance in each class to enhance the performance of the classifier. In the future, this study still needs more improvements. It can be done by adding more features, such as the distance from the epicenter to the shoreline, to increase the performance of the classifier.

REFERENCES

[1] Lessy, M. The Earthquake Zoning in Indonesia, http://www.academia.edu/4517794/Zonasi_Gempa_bumi_di_Indonesia *accessed on 3 March 2019*.
[2] The Official Website of Meteorology and Geophysics Body of Indonesia, http://www.bmkg.go.id, *accessed on 3 March 2019.*
[3] Suhardjono, et al., Tsunami early warning for broadcasting institutions in Indonesia, Information Guidebook, 2012
[4] Indonesia Tsunami Early Warning System, http://inatews.bmkg.go.id, *accessed on 1 January 2019*.
[5] The Official Website of National Oceanic and Atmospheric Administration https://www.noaa.gov/, *accessed on 20 September 2018*.
[6] T. Shibayama, M. Esteban, I. Nistor, et al. Classification of Tsunami and Evacuation Areas, Nat Hazards (2013) 67: 365. https://doi.org/10.1007/s11069-013-0567-4
[7] Oktariadi, O. 2009. Penentuan Peringkat Bahaya Tsunami dengan Metode Analytical Hierarchy Process (Studi kasus: Wilayah Pesisir Kabupaten Sukabumi) [Determination of Tsunami Hazard Rating by Analytical Hierarchy Process Method (Case study: Coastal Area of Sukabumi Regency)], Jurnal *Geologi Indonesia*, 4(2): 103-116.
[8] Abidin, Z. 2011. Implementasi Algoritma C 4.5 Untuk Menentukan Tingkat Bahaya Tsunami. *Seminar Nasional Informatika 2011* [Implementation of C 4.5 Algorithm for Determining Tsunami Hazard Levels. National Informatics Seminar 2011], UPN Veteran Yogyakarta, 2 July 2011.
[9] Dilectin, H.D., Mercy, R.B.V. 2012. Classification and dynamic class detection of real-time data for tsunami warning system, *International Conference on Recent Advances in Computing and Software Systems (RACSS),* pp 124,129.
[10] Hadi, S., Nursantika, and D. Purwanti, I. 2012. Integrating computer vision technique to support Tsunami Early Warning.
[11] J. Wachter, et.al. Development of tsunami early warning systems and future challenges, Natural Hazard and Earth System Sciences, 2012.
[12] P. Kumar, P. Roy, D. Dogra, Independent Bayesian classifier combination-based sign language recognition using facial expression, Information Sciences, Volume 428, 2018.

[13] Michael. S. 2002. Better Bayesian filtering, Meat Slicer: Spam Classification with Naive Bayes and Smart Heuristics. Downloaded at http://vvww.paulgraham.com/better.html

[14] Romer, H., et al. 2012. Potential of remote sensing techniques for tsunami hazard and vulnerability analysis – a case study from Phang-Nga province, Thailand. Natural Hazar and System Sciences. 12, pp. 2103–2126. doi:10.5194/nhess-12-2103-2012

[15] Sambah, A.B., Miura, F. Integration of Spatial Analysis for Tsunami Inundation and Impact Assessment. Journal of Geographic Information System, 2014, 6, 11-22.

Tsunami Early Warning Detection using Bayesian Classifier

ORIGINALITY REPORT

11%	10%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ docshare.tips

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		